Coincidence Complex Networks

Luciano da Fontoura Costa
luciano@ifsc.usp.br

São Carlos Institute of Physics – DFCM/USP

27rd Oct 2021

Abstract

Complex networks, which are the main object of the relatively recent area of network science, have been wide and extensively used for representing, characterizing, and modeling an ample range of structures and phenomena from both theoretical and applied perspectives. The present work describes the application of the recently introduced real-valued Jaccard and coincidence similarity indices for building complex networks in which two nodes are linked whenever their similarity is greater than a given threshold. Weighted networks can also be obtained by taking the similarity indices as weights. It is shown that the proposed approach allows enhanced performance when compared to the Pearson correlation and cosine distance-based approaches. The reported methodology and results paves the way to a significant number of possible theoretical and applied developments.

Invisibly linked, the universe realizes its complexity.

LdaFC

1 Introduction

Though relatively recent, the area of network science ([1, 2, 3]) has established itself not only from the perspective of theoretical contributions, but also as a consequence of its ample range of applications to the most diverse areas and problems [4]. Basically, a complex network is a graph presenting structure significantly distinct from other simpler, more regular, counterparts as far as a relatively comprehensive set of respective topological measurements are concerned [5].

The ability of complex networks, which constitutes network science's main object of study, to represent virtually every discrete systems has contributed significantly to the success of this area. This has allowed, among many other possibilities, complex networks to be employed to represent several complex structures including but by no means limited to the Internet and WWW, transportation systems including highways and airports, the structure of cities, protein interaction, social and scientific networks, as well as signals and images.

The representation of a structure or system as a complex network demands the adoption of some approach for defining the network nodes, and then the application of some objective criterion for linking nodes. Actually, there seems to be intrinsic complementary relationships between connectivity, similarities and features [6].

Several possibilities have been considered in the literature but, basically, they can be organized into two main classes of approaches, namely *distance-based* and *similarity-based* criteria.

Given that linking of two nodes is typically expected to reflect shared properties of them, similarity-based criteria for deriving respective networks represent a suitable choice. Often employed approaches include the consideration of some similarity measurement such as the Pearson correlation coefficient or cosine similarity, as well as set-based alternatives, such as the Jaccard index (e.g. [7, 8]).

Indeed, the choice of the similarity measurement to be adopte depends intrinsically on the nature of the data respective to the system to be represented. In case of categorical data, approaches such as the Jaccard index provide an interesting choice. On the other hand, in case the data involves integer or real-valued quantities, numeric indices such as the cosine similarity and Pearson correlation coefficient have been frequently employed.

More recently, the Jaccard index has been generalized to real-valued data [7, 8] through the multiset approach (e.g. [9, 10, 11, 12, 13, 14]). Though a version of the Jaccard similarity index adapted for multisets has been considered for some time, it was not able to take into account the more general situation of having negative data values. This further generalized version of the Jaccard index has been called the *real-valued Jaccard* similarity

index.

In addition, the classic Jaccard index has been shown not to be able to take into account the relative interiority of the two compared sets, which motivated the proposal [7, 15] of the *coincidence index*, which has been understood as the product of the interiority (or homogeneity) and real-valued Jaccard indices.

Recent results [7, 15] have indicated that product-based similarity indices, which include both the cosine similarity and Pearson correlation coefficient, tend to be relatively more tolerant to the similarity than indices based on combination of the minimum and maximum operations, such as the Jaccard index. Actually, it has been shown [15] that all these similarity approaches are ultimately related to the Kronecker's delta function, which can be understood as the prototypical approach to quantifying the similarity between two values, though in a quite strict manner.

Therefore, the real-valued Jaccard and coincidence indices, both of which relying on maximum and minimum operations, constitute an interesting alternative to be considered in situations where the cosine, Pearson correlation, and classic Jaccard have been traditionally adopted. In fact, the fact that these two indices reflect the sign relationship between the compared values has also allowed them to be considered as indications os statistical joint variations [7].

The present work aims at studying the potential of the real-valued Jaccard and coincidence similarity index as the basis for mapping datasets, involving one or more features corresponding to real values, into respective complex networks.

Though undirected complex networks are henceforth assumed in this work, the reported concepts and methods can be readily extended to directed comples network research.

2 Basic Concepts

A complex network is basically a graph, being therefore composed of N nodes or vertices and E connections, edges, or links. Complex networks are often represented in terms of the respective adjacency matrices or lists of edges. The former of these possible representations involves using a matrix A where a connection from a node i to a node j implies $A[i,j] = 1$, with $A[i,j] = 0$ being otherwise enforced.

Given a network represented in terms of its respective adjacency matrix, several topological measurements (e.g. [3]) can be obtained, some of which are adopted in the present work and reviewed as follows.

The *degree* of a node i corresponds to the respective number of connections established with that node. It can be readily calculated as:

$$k_i = \sum_{j=1}^{N} A[i,j] \tag{1}$$

The Euclidean distance between two feature vectors $\vec{x}$ and $\vec{y}$ with N components can be expressed as:

$$d_E(\vec{x},\vec{y}) = \sum_{i=1}^{N} \sqrt{(x_i - y_i)^2} \tag{2}$$

A possibly induced similarity can be defined as:

$$s_E(\vec{x},\vec{y}) = \frac{1}{d_{E,n}(\vec{x},\vec{y})} \tag{3}$$

Another frequently adopted approach, namely the cosine distance, relies on the inner product between two sets $\vec{x}$ and $\vec{y}$, which can be expressed as:

$$\langle \vec{x},\vec{y} \rangle = \sum_{i=1}^{N} x_i \; y_i = |\vec{x}||\vec{y}| \cos(\theta) \tag{4}$$

In order to avoid biases typically implied by varying magnitudes of the several features, the statistic linear transformation of *standardization* is frequently adopted. This transformation is expressed as follows:

$$\tilde{f}_i = \frac{f_i - \mu_i}{\sigma_i} \tag{5}$$

It can be shown that each of the new features will have zero means and unit standard deviation. In addition, most of the feature values become comprised within the interval $[-2, 2]$, though this depends o their respective statistical distributions. In addition, the newly obtained variables become non-dimensional.

Observe that the application of the standardization procedure intrinsically implies that any index or measurement applied to quantify the similarity between data elements will need to be able to cope with negative values, which is not the case of the classic Jaccard index, or even its standard multiset generalization to non-negative values.

3 Real-Valued Jaccard and Coincidence Similarity Indices

The classic Jaccard index [7]) for quantifying the similarity between two sets A and B is defined as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

In a recent generalization [7, 15], henceforth referred to as *real-valued Jaccard similarity index* the Jaccard index

has been extended to take into account possibly negative, real data values [7, 8, 15]. The real-valued Jaccard similarity index can be expressed as:

$$s_1(A,B) = \frac{\sum_{i\in S} s_{x_i y_i} \min\{s_{x_i}x_i, s_{y_i}y_i\}}{\sum_{i\in S} \max\{s_{x_i}x_i, s_{y_i}y_i\}} \quad (7)$$

where x_i and y_i correspond to the multiplicities of the sets A and B, and S is the support of the multisets, which generally corresponds to the union of the elements in A and B.

The classic Jaccard index, as well as its generalization to multisets and real-values, have been shown not be able to take into account the relative interiority of the two sets [7].

The relative interiority (or homogeneity) between any two multisets can be quantified as:

$$I(A,B) = \frac{\sum_{i\in S} \min\{s_{x_i}x_i, s_{y_i}y_i\}}{\min\left\{\sum_{i\in S} s_{x_i}x_i,\ \sum_{i\in S} s_{y_i}y_i\right\}} \quad (8)$$

$$(9)$$

This measurement, however, does not take into account how much of the two multisets are incorporated into the intersected region, a property that is effectively reflected in the Jaccard index.

In oder to combine the two measurements, both of which relating directly to the concept of similarity between two sets, the *coincidence index* between any two multisets A and B has been proposed [7, 8] as corresponding to the product of the respective interiority and Jaccard indices. Therefore, we have that the coincidence index can be calculated as:

$$C(A,B) = I(A,B)J(A,B) \quad (10)$$

4 Building Coincidence Networks

In this work, we will approach the situation in which a set of N data observations, saamples, element or individuals, have been priorly characterized in terms of M respective measurements. This can be understood as a table or matrix where each of the N rows corresponds to a data observation, while the columns contain the respective measurements or features x_i, $i = 1, 2, \ldots, M$. Experimental data are normally represented and stored in this format.

Given such a table, we aim at obtaining a respective representation as a complex network. The commonly adopted procedure consists of understanding each data element as a node, which links are established based on the similarity (or difference) between the respective features.

Frequently applied methodologies rely direct or indirectly on the Pearson correlation coefficient between the involved features. Here, we adopt the recently introduced real-valued Jaccard and coincidence indices.

5 The Iris Dataset

For simplicity's sake, we consider the iris dataset (e.g. [16]), which has been extensively considered in pattern recognition and machine learning. This dataset contains 50 samples of each of 3 iris flower species, namely *Iris setosa*, *Iris virginica* and *Iris versicolor*. Each of the data elements is characterized in terms of four respective real-valued features.

Figure 1 depicts the iris data set considering the second and third original features. All four real-valued original features are considered in all cases herein. In addition, in order to avoid effects of the varying magnitudes of the four original measurements, which could therefore bias the network representations, we consider the respectively standardized dataset.

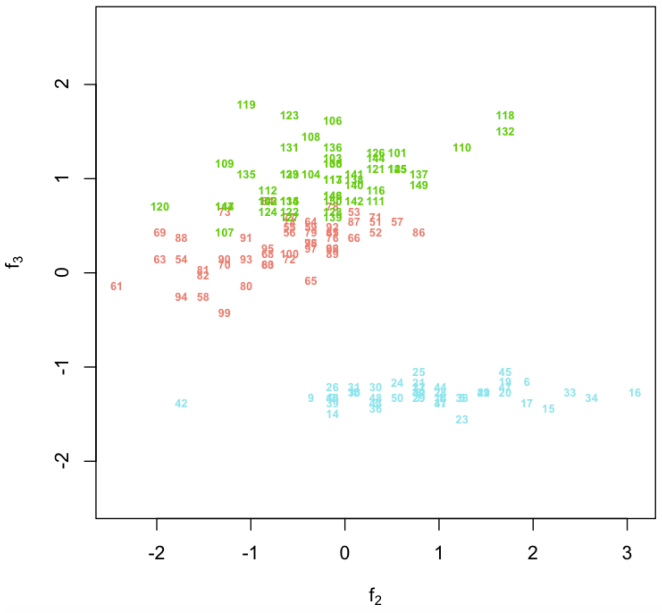

Figure 1: Scatterplot of the Iris dataset considering original features 2 and 3. Each of the three iris species have been shown in specific respective colors. The numbering is as in the original dataset, with 50 individuals of each species number subsequently from 1 to 50, 51 to 100, and 101 to 150. One group is markedly separated, but the other two are adjacent and even present some overlap.

The overall structure is, however, difficult to be observed because it involves a four-dimensional space. Nevertheless, we can see from Figure 1 that this dataset contains a well separated cluster (in blue), while the other two groups are adjacent and even present some overlap. It has particularly difficult to separate these two clusters in the literature.

The application of the Euclidean similarity to obtain a respective complex network resulted in the structure shown in Figure 2 .

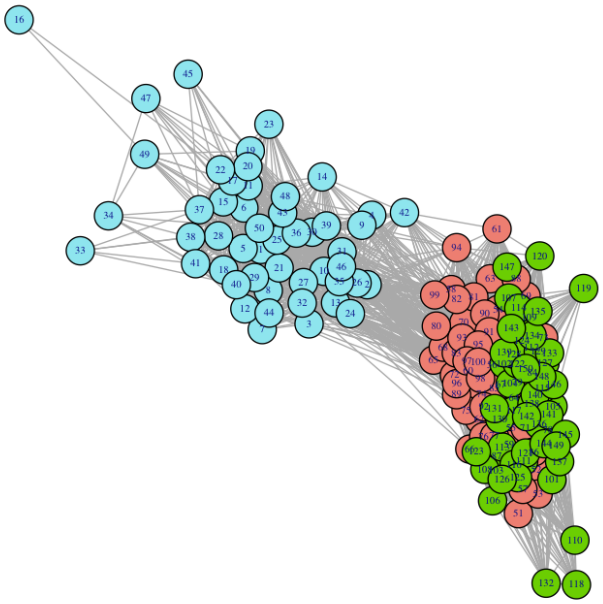

Figure 2: Complex network of the iris dataset obtained while considering the *Euclidean similarity* and all four original features. This result resembles the relative distribution of points in the original dataset.

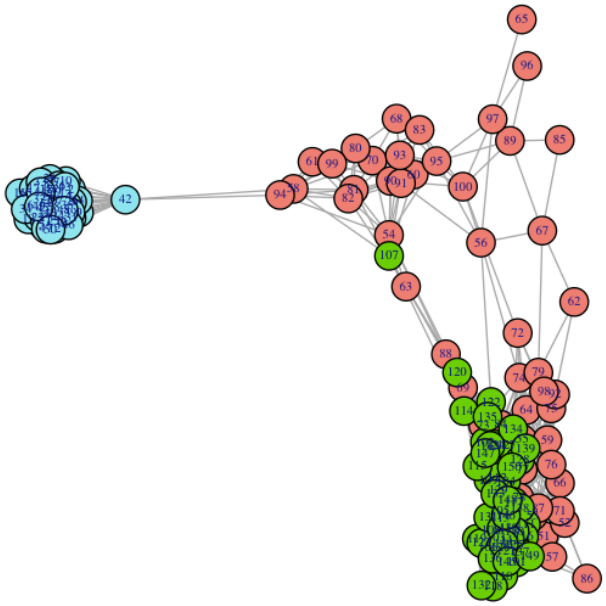

Figure 3: Complex network of the iris dataset obtained by using the *real-value Jaccard index* and the four original features. Several remarkable characteristics can be identified in this result. First, we have that the blue group, which was already well separated in the original feature space, coalesced into a very compact community. Second, we have that a substantial portion of the salmon group resulted relatively well separated from the other two groups. In addition, the obtained enhanced modularity indicates the improved accuracy and discriminability allowed by this approach.

The obtained complex network resembles closely the relative position and separation of the original dataset as projected onto the second and third original measurements shown in Figure 1.

Now, we proceed to applying the recently introduced real-valued Jaccard index to the iris dataset, yielding the results depicted in Figure 1.

These results reveal a surprising separation of the groups. Not only the blue group has coalesced into a very compact community, but a substantial portion of the salmon group has also resulted separated from the other two groups. In addition, this result allows a substantially more detailed representation of the interrelationship between the involved data elements, as well as the underlying modular structure. Observe also that the outliers points in Figure 1 have been duly incorporated at the border of the obtained network.

In a sense, the adoption of the real-valued Jaccard similarity acted almost as a community detection algorithm, strongly emphasizing the respective modular structure and centrality of the nodes.

In order to illustrate the interiority similarity, we obtained the respective complex network shown in Figure 1.

These results well illustrate that, despite the effectivity of the interiority index in quantifying the relative interiority of a pair of sets, this index cannot actually provide a more complete picture of the similarity between the several involved data elements.

Figure 5 shows the network obtained by adopting the coincidence index.

The already identified enhanced discriminative potential of the coincidence index [8, 17] has allowed a network representation of the iris dataset that is even more detailed and structured regarding the interrelationship between the data elements, as well as the overall data modular structure. In addition, the three groups became even better separated, with the green and salmon group noa presenting an interstice region.

6 Concluding Remarks

Network science has consolidated itself as an important area, with many theoretical and applied contributions and

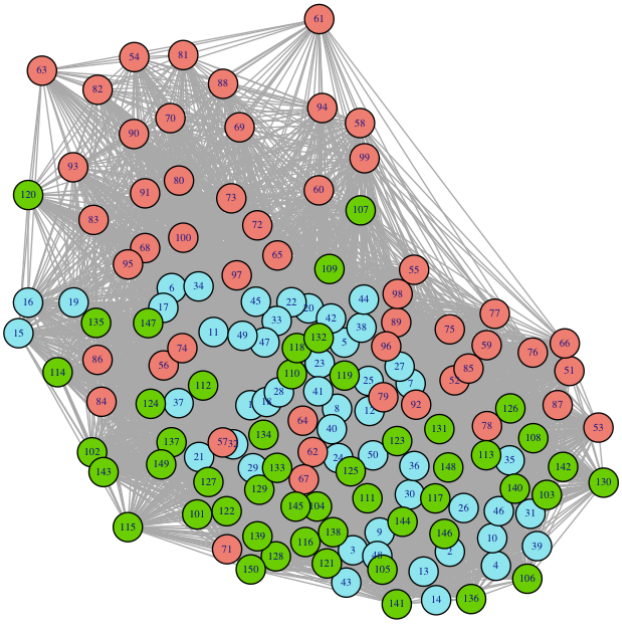

Figure 4: Complex network of the iris dataset obtained from the *interiority similarity* between each pair of original data elements. The obtained results corroborates that, despite the ability of the interiority index to quantify the relative interiority between two msets, this index does not actually take into account the relative intersection are with respect to the size of the msets. Observe that, by mset, we are meaning the mset of features associated to each data individual.

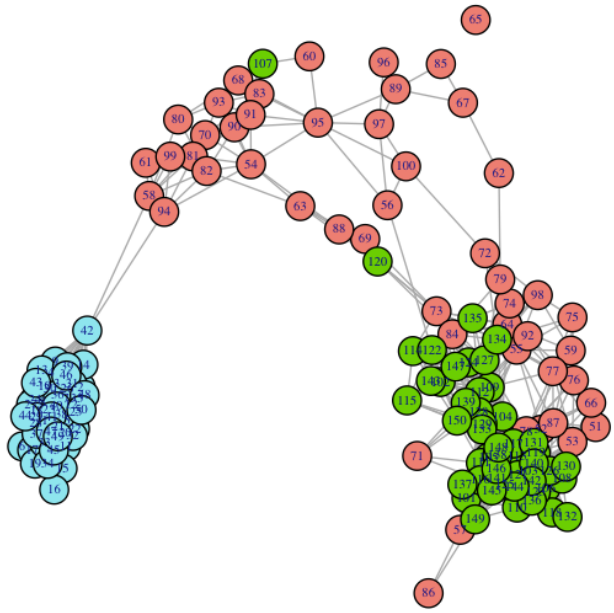

Figure 5: Complex network of the iris dataset obtained by using the *coincidence index* and the four original features. Though resembling the results obtained for the real-valued Jaccard index, this results can be observed to provide even greater potential for revealing the intricacies of the modular structure of the iris dataset. Observe that, as a consequence, the green and salmon groups resulted even more separated than in the case of the real-valued Jaccard index.

enduring potential. Much of these success stems from the ability of complex networks to *represent* virtually every possible discrete structure or system, while continuous counterparts can always be discretized to a given resolution.

One of the enduring challenges in network science concerns how to transform datasets into networks. Though the Euclidean distance, Pearson correlation coefficient, and cosine similarity have been ubiquitously applied, other measurements based on the minimum and maximum binary operations have not been mostly limited to categorical or binary data.

The present work set out at addressing the possibility of using the recently introduced real-valued Jaccard and coincidence indices as the basis for deriving complex network representations of real-valued data, therefore also encompassing other types of data such as integer and categorical. Actually, an immediate benefit of such an approach constitutes in its being inherently capable of taking into account hybrid datasets involving several types of features, which is allowed by the fact that both the real-valued Jaccard and coincidence index can immediately adapt to these other types of data.

The results obtained with respect to the well-known iris dataset revealed some remarkable findings. First, we have that the enhanced precision and discrimination potential of the real-valued and Jaccard indices, already observed for experimental [17] and theoretically [15], allowed complex networks to be obtained which reveal to a great level of detail the intrinsic relationships between the data elements as well as the overall data modular structure. Comparatively, the Euclidean similarity-based obtained representation can be understood as a direct translation of the original pairwise distances between the original datasets, without revealing any additional structure or properties of the data elements interrelationship. In a sense, the real-valued Jaccard and coincidence indices action resulted effectively in a community detection methodology.

The proposed methodology bears significant potential for ample application in complex network research, allowing a unified and effective manner of obtaining representations of the most diverse type of data and features, including cases of mixed types of measurements.

Acknowledgments.

Luciano da F. Costa thanks CNPq (grant no. 307085/2018-0) and FAPESP (grant 15/22308-2).

References

[1] A.L. Barabási and Pósfai M. *ience*. Cambridge University Press, 2016.

[2] M. Newman. *Networks: An introduction*. Oxford University Press, 2010.

[3] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.

[4] L. da F. Costa, O.N. Oliveira Jr., G. Travieso, F.A. Rodrigues, P.R. Villas Boas, L. Antiqueira, M.P. Viana, and L.E.C. Rocha. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, 60(3):329–412, 2011.

[5] L. da F. Costa. What is a complex network? https://www.researchgate.net/publication/324312765_What_is_a_Complex_Network_CDT-2, 2018. [Online; accessed 05-May-2019].

[6] E. K. Tokuda, C. H. Comin, and L. da F. Costa. Complex systems: Features, similarity and connectivity. *Physics Reports*, 861:1–41, 2020.

[7] L. da F. Costa. Further generalizations of the Jaccard index. https://www.researchgate.net/publication/355381945_Further_Generalizations_of_the_Jaccard_Index, 2021. [Online; accessed 21-Aug-2021].

[8] L. da F. Costa. Multisets. https://www.researchgate.net/publication/355437006_Multisets, 2021. [Online; accessed 21-Aug-2021].

[9] J. Hein. *Discrete Mathematics*. Jones & Bartlett Pub., 2003.

[10] D. E. Knuth. *The Art of Computing*. Addison Wesley, 1998.

[11] W. D. Blizard. Multiset theory. *Notre Dame Journal of Formal Logic*, 30:36—66, 1989.

[12] W. D. Blizard. The development of multiset theory. *Modern Logic*, 4:319–352, 1991.

[13] P. M. Mahalakshmi and P. Thangavelu. Properties of multisets. *International Journal of Innovative Technology and Exploring Engineering*, 8:1–4, 2019.

[14] D. Singh, M. Ibrahim, T. Yohana, and J. N. Singh. Complementation in multiset theory. *International Mathematical Forum*, 38:1877–1884, 2011.

[15] L. da F. Costa. On similarity. https://www.researchgate.net/publication/355792673_On_Similarity, 2021. [Online; accessed 21-Aug-2021].

[16] Wikipedia. Iris flower data set. https://en.wikipedia.org/wiki/Iris_flower_data_set, 2021. [Online; accessed 30-Oct-2021].

[17] L. da F. Costa. Comparing cross correlation-based similarities. https://www.researchgate.net/publication/355546016_Comparing_Cross_Correlation-Based_Similarities, 2021. [Online; accessed 21-Oct-2021].